Alzheimer's Research & Therapy

REVIEW

Open Access

Deep learning-based speech analysis for Alzheimer's disease detection: a literature review

Qin Yang[2], Xin Li[1,2*], Xinyun Ding[2], Feiyang Xu[2] and Zhenhua Ling[1]

Abstract

Background: Alzheimer's disease has become one of the most common neurodegenerative diseases worldwide, which seriously affects the health of the elderly. Early detection and intervention are the most effective prevention methods currently. Compared with traditional detection methods such as traditional scale tests, electroencephalograms, and magnetic resonance imaging, speech analysis is more convenient for automatic large-scale Alzheimer's disease detection and has attracted extensive attention from researchers. In particular, deep learning-based speech analysis and language processing techniques for Alzheimer's disease detection have been studied and achieved impressive results.

Methods: To integrate the latest research progresses, hundreds of relevant papers from ACM, DBLP, IEEE, PubMed, Scopus, Web of Science electronic databases, and other sources were retrieved. We used these keywords for paper search: (Alzheimer OR dementia OR cognitive impairment) AND (speech OR voice OR audio) AND (deep learning OR neural network).

Conclusions: Fifty-two papers were finally retained after screening. We reviewed and presented the speech databases, deep learning methods, and model performances of these studies. In the end, we pointed out the mainstreams and limitations in the current studies and provided a direction for future research.

Keywords: Alzheimer's disease detection, Speech analysis, Deep learning

Background

Dementia is the most common neurodegenerative disease among the elderly, of which Alzheimer's disease (AD) is the most common type. According to data from the World Health Organization, the current incidence of AD has shown a significant upward trend in recent years, and the number of patients will reach 152 million in 2050 [1], which will affect the health of the people seriously and cause an enormous economic burden on home care and social security. However, effective treatment is not yet available. Studies have shown that the early diagnosis and intervention based on early assessment and screening of cognitive impairment can help maintain healthy brain activity, retard irreversible brain decline, delay disease progression, and prolong patient life [2]. In this case, early detection of mild cognitive impairment (MCI), which is the early stage of AD, is very important for delaying cognitive state decline.

Currently, the mainstream clinical methods for AD detection include scale testing, brain magnetic resonance imaging measurement (MRI), cerebrospinal fluid analysis, etc. These methods are either time-consuming and labor-intensive, or expensive and unfriendly to subjects' experience. In general, traditional AD detection methods such as magnetic resonance imaging, positron

*Correspondence: leexin@ustc.edu.cn

[2] iFlytek Research, iFlytek Co.Ltd, Hefei, China
Full list of author information is available at the end of the article

emission tomography (PET) imaging, and cerebrospinal fluid (CSF) assays [3], are not appropriate for large-scale nationwide early AD screening applications. Therefore, some studies focus on developing a cheaper and more convenient method to detect AD.

Relevant studies have shown that language disorders usually appear in the early process of AD, and it is possible to detect AD by capturing the acoustic and linguistic features of subjects through audio and automatic speech recognition technology [4–6]. Some studies have given the results of studies on distinguishing characteristics between AD and healthy control (HC) group. Compared with cognitive normal people, AD patients usually speak more slowly with more pauses between words [7] and suffer from word finding and word retrieval difficulties [6, 8, 9].

Dozens of speech-based methods have been explored for the research on AD detection. Studies have shown that the acoustic measures have a high correlation with pathological language features and voice changes in automatic language processing were proven to be useful for AD detection [10, 11]. In addition, previous studies of speech pathology have revealed that people with dementia have linguistic manifestations including pauses, filler words, restarts, repetitions, and incomplete statements. Fraser, K.C. et al. extracted linguistic features such as semantics, syntax, and information and achieved 91% accuracy [4] in the AD detection task by using the logistic regression classifier. Liu, Z. et al. extracted and fused duration features, acoustic features, linguistic features, the AD detection, and linguistic features, and finally obtained 81.9% accuracy of AD detection based on the logistic regression classification method [12]. In addition to these, Satt, A. et al. utilized recordings while subjects completed cognitive tasks to extract relevant acoustic features, and achieved an accuracy of 87% in the classification between AD and control [5].

With the wide application of deep learning, we can find that neural networks have made significant progress in the field of speech modeling. Hinton, G. et al. applied deep neural networks (DNNs) to acoustic modeling and obtained better recognition results than Gaussian Mixed Model (GMM), thus opening up a new field in speech recognition [13]. Therefore, researchers began to try to apply various deep learning methods to the field of speech-based AD detection. Rosas, D.S. et al. extracted linguistic features and used a 3-layer neural network reaching a binary classification accuracy of 78.3% [14]. However, there is fewer speech data for Alzheimer's patients, and the improvement in classification results is relatively small by using neural networks. Recent studies have shown that pre-trained models such as BERT [15] achieve promising results on a variety of benchmark tasks, and can capture a wide range of linguistic facts including lexical knowledge, phonology, syntax, semantics, and pragmatics without a lot of data. Apart from this, the pre-trained automatic speech recognition (ASR) model can not only get the transcribed text of speech but also extract acoustic embeddings which can be used to represent the conversion in speech for better automatic analysis. Toth, L. et al. obtained phonetic segmentation and label of the input signal by applying an ASR model based on a special convolutional deep neural network, thereby obtaining acoustic features such as speech rate, pause, and hesitation rate [16]. Judging by the current research trends, the deep learning method is the most mainstream method for AD detection now.

Simultaneously, some review papers on AD detection have also been published, such as a systematic review about speech-based detection and classification of AD written by Inês Vigo et al. [17]. However, most of the classification methods are based on traditional machine learning methods, which have certain limitations due to the excellent performance achieved by deep learning methods in AD detection.

Therefore, this paper focuses on deep learning-based speech analysis for AD detection. This research paper is organized as follows: the objects of this review in the "Objectives" section, the search and selection process is introduced in the "Materials and methods" section, the results in the "Results" section, the discussion of these selected papers in the "Conclusions" section, and the limitation of our work and our future goals in the "Discussions" section.

Objectives

To make a comprehensive discussion on the current application of deep learning in speech-based AD detection, this review conducted a systematic analysis of selected papers in response to the following 5 questions:

(1) What were the characteristics of the databases involved in reported studies?
(2) What deep learning model architectures were included in reported studies?
(3) How were these deep learning model architectures used in reported studies?
(4) What classification performance has been achieved?
(5) What were the mainstreams and limitations of reported studies?

Materials and methods

Search process

Our searches were conducted on the following electronic databases: ACM, DBLP, IEEE, PubMed, Scopus, and Web

of Science. Unlike most previous review papers on "Alzheimer's disease detection" [18], we paid more attention to these papers which used deep learning methods to analyze speech data of elderly people in different health states (AD, MCI, and HC). Therefore, we used the following keywords for paper search: (Alzheimer OR dementia OR cognitive impairment) AND (speech OR voice OR audio) AND (deep learning OR neural network). Figure 1 listed all the search strategies. The last search was conducted on 19 January 2022.

Selection process

The exclusion criteria were as follows: (1) studies that did not use deep learning methods; (2) studies do not focus on speech or text data; (3) studies without a group of MCI and AD; (4) papers that were not written in English; (5) studies cannot find the full text. Initial study selection was performed by two reviewers independently. To minimize the bias in selecting studies, papers that were not sure to include were resolved in a discussion with the third reviewer.

Data extraction and synthesis

The analyzed data in our studies include database names, task types, language types, label distributions, and whether the databases include an audio or corresponding transcript or not.

Results

Study selection

The detail of our search process is displayed in Fig. 2 through a flow diagram. other source papers retrieved from the ADReSS website [19] which were not found in the other six sources. After the search process, a total of 710 papers were retrieved; 293 duplicates were removed by Endnote and manual screening. After screening by our exclusion rules, 52 studies were finally included.

Speech databases

In the process of building a deep learning model, a high-quality database can improve the quality of model training and the accuracy of prediction. At present, several speech databases for cognitive impairment of the elderly have been established around the world, providing great support for researchers to explore more efficient cognitive impairment assessments.

According to our research, the linguistic tasks involved in the studies on AD detection based on deep learning methods can be divided into three categories: semantic verbal fluency (SVF), spontaneous speech (SS), and reading. Therefore, the related speech databases will also be introduced from these three aspects in this review (Table 1).

Semantic verbal fluency tasks

The semantic verbal fluency (SVF) test has high sensitivity and specificity for the diagnosis of AD, so it is widely

Search Strategy

ACM
[[Full Text: alzheimer] OR [Full Text: dementia] OR [Full Text: "cognitive impairment"]] AND [[Full Text: speech] OR [Full Text: voice] OR [Full Text: audio]] AND [[All: "deep learning"] OR [All: "neural network"]]

DBLP
alzheimer|dementia|"cognitive impairment" "speech"|"voice"|"audio"

IEEE
("All Metadata":dementia OR "All Metadata":alzheimer OR "All Metadata":cognitive impairment) AND ("All Metadata":speech OR "All Metadata":audio OR "All Metadata":voice) AND ("All Metadata":deep learning OR "All Metadata":neural network)

PubMed
(dementia OR alzheimer OR "cognitive impairment") AND (speech OR audio OR voice) AND ("deep learning" OR "neural network")

Scopus
(TITLE-ABS-KEY (dementia OR alzheimer OR "cognitive impairment") AND TITLE-ABS-KEY (speech OR audio OR voice) AND TITLE-ABS-KEY ("deep learning" OR "neural network"))

Web of Science
(dementia OR alzheimer OR "cognitive impairment") AND (speech OR audio OR voice) AND ("deep learning" OR "neural network")

Fig. 1 Search strategies in ACM, DBLP, IEEE, PubMed, Scopus, and Web of Science

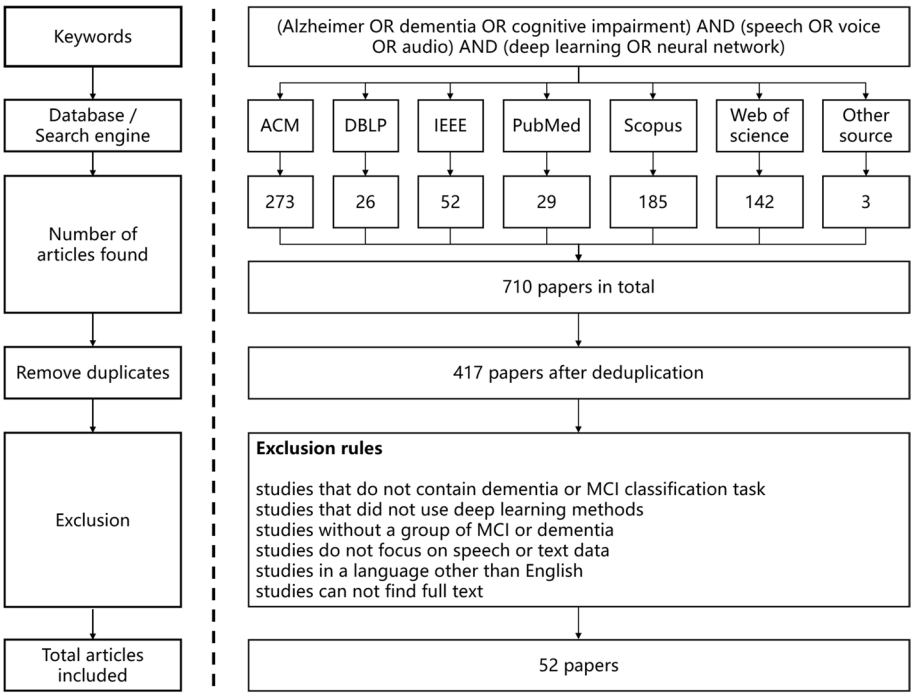

Fig. 2 Flow diagram of the search and selection process in our study

used to assess language skills, semantic memory, and executive functions of AD patients. During the SVF task, patients were asked to list all names they can remember from a category within one minute, such as animals, vegetables, and locations [20].

Animal naming

The subjects were asked to say the name of the animal they can think of as quickly as possible within 60 s and were reminded if they stop. At the end of the 60 s, the total number of animals (NOT including repetitions or non-animal words) were counted as their scores [22].

Lopez-De-Ipina, K. et al. constructed a well-distributed animal naming database called PGA-OREKA, which presents a novel proposal based on automatic analysis of speech and disfluencies aimed at supporting MCI diagnosis [21]. The PGA-OREKA database contains 62 healthy people and 38 MCI patients, and it is a subset of the cohort of the Gipuzkoa-Alzheimer Project (PGA) of the CITA-Alzheimer Foundation which includes 187 healthy people and 38 MCI patients.

Vegetable and location naming

Similar to animal naming, in vegetable and location naming tests, subjects were asked to say as many words related to the designated topic as possible within one minute. Chien, Y.W. et al. from National Taiwan University constructed a fluency test database based on the Mandarin_Lu corpus [23]. Mandarin_Lu corpus from DementiaBank contains interview recordings of 52 AD patients [24], Chien, Y.W. et al. selected 30 patients and segmented the first-minute response of the audio data, and then recruited 30 additional healthy subjects to complete vegetable and location naming tasks.

Spontaneous speech tasks

Spontaneous speech (SS) means speech without responding to a question. Temporal parameters of spontaneous speech have been proven to be able to provide sensitive measures of a subject's speech and language skills [25]. Several different types of spontaneous language tasks are covered in this review paper: conversation/interview speech, event description, recall story, and picture description.

Conversation/interview speech

Through natural language processing and analysis of the subject's speech obtained from free and simple conversational speech, some vital biological features that reflect early signs of AD can be extracted for early screening.

Lopez-De-Ipina, K. et al. built up a multicultural and multilingual database called AZTIAHO [26], which contains 20 h of video recordings of 50 healthy control and 20 AD patients. The recordings consisted of conversational

Table 1 Dementia-related speech-based databases information

Task name			Database name	Abbreviation	Language	Label distribution	Speech included	Transcription included
VF	SVF	Animal naming	PGA-OREKA [20]	VF1	Spain	HC (62)/MCI (38)	Yes	No
			- [21]	VF2	French, Dutch, and German	HC (66)/MCI (66)	Yes	-
		Vegetable naming	Mandarin_Lu (DementiaBank) + NTU dataset [22]	VF3	Chinese	AD (30)/HC (30)	Yes	No
		Location naming	Mandarin_Lu (DementiaBank) + NTU dataset [22]	VF4	Chinese	AD (30)/HC (30)	Yes	No
SS	Conversation/interview		PROMPT database [23]	cvs1	Japanese	Dementia (49)/MCI(42)/ HC(72)	Yes	-
			- [24]	cvs2	Italian	eD (16)/MCI (32)/HC (48)	Yes	Yes
			The Carolina Corpus Conversation database [14]	cvs3	English	AD (30)/HC (16)	Yes	Yes
			IVA dataset [25]	cvs4	English	ND (21)/MCI (24)/HC (25)	Yes	-
			The Hungarian MCI-mAD Database [26, 27]	cvs5	Hungarian	MCI (48)/HC (36) AD (25)/MCI (25)/HC (25)	Yes	No
			AZTIAHO AZTIAHORE [28]	cvs6	multilingual	AD (20)/HC (50) AD (20)/MCI (20)	Yes	No
			- [29]	cvs7	Italian	MCI (19)/HC (20)	Yes	Yes
			- [16]	cvs8	Hungarian	MCI (32)/HC (19)	Yes	No
			Framingham Heart Study Dataset [30]	cvs9	English	Dementia (223)/MCI (309)/ HC (291)	Yes	No
	Recall		NTUHV dataset [31]	recall1	Chinese Taiwan	AD (40)/HC (40) MCI (30)/HC (30)	-	-
			The Wallet Story from ABCD [32]	recall2	Brazilian Portuguese	MCI (23)/HC (12)	-	-
			The Lucia Story Datasets from BALE [32]	recall3	-	AD (9)/HC (80)	-	-
			The Hungarian MCI-mAD Database [26, 27]	recall4	Hungarian	MCI (48)/HC (36) AD (25)/MCI (25)/HC (25)	Yes	No
			- [33]	recall5	Chinese Taiwan	HC (30)/AD (30)	Yes	No
	PD	Cookie theft	pitt Corpus [34]	SS-PD-CT1	English	HC (244)/MCI (309)	Yes	Yes
			ADReSS [19]	SS-PD-CT2	English	AD (78)/non-AD (78)	Yes	Yes
			ADReSSo [35]	SS-PD-CT3	English	AD (87)/HC (79)	Yes	Yes
			Wisconsin Longitudinal Study (WLS) [36]	SS-PD-CT4	English	AD (115)/HC (839)		
			- [37]	SS-PD-CT5	English	AD (26)/HC (46)	Yes	Yes
			NTUHV dataset [31]	SS-PD-CT6	Chinese Taiwan	40 AD/40 HC//30HC/30 MCI	-	-
		-	- [29]	SS-PD1	Italian	MCI (19)/HC (20)	Yes	-
		-	- [24]	SS-PD2	Italian	eD (16)/MCI (32)/HC (48)	Yes	Yes
		MINI-PGA	MINI-PGA [12]	SS-PD3	Spain	AD (6)/HC (12)	Yes	No
		The Dog Story	The Dog Story from BALE [32]	SS-PD4	Brazilian Portuguese	AD (12)/MCI (12)/HC (82)	No	Yes
		The Cinderella Dataset	The Cinderella Dataset [32]	SS-PD5	Brazilian Portuguese	AD (20)/amnestic MCI(20)/ HC (20)	Yes	Yes
Reading	Transcripts Reading		Gothenburg MCI study [38]	Reading	Swedish	AD (25)/HC (30)	Yes	No

The full names of abbreviations can be found in "Abbreviations"

speech where subjects tell pleasant stories or feelings and interact with each other.

Day/life/dream description

During these tests, subjects were asked to spontaneously describe events such as tell about the day yesterday in detail. Gosztolya, G. et al. established the Hungarian MCI-mAD Database [27], which recorded 225 voices of 75 subjects (25 AD, 25 MCI, and 25 HC).

Recall story

Subjects were given orally presented stories, reading materials, or films to learn the specific stories. Then they were asked to recall and retell the story spontaneously twice, immediately and in a few minutes, to the examiners without reference to those materials.

The Wallet Story database was collected based on the immediate and late retelling of a memorized story from (Bayles and Tomoeda, 1993), which is the evaluation of the episodic memory, one of a standardized test battery named ABCD (Arizona Battery for Communication Disorders) for the comprehensive assessment and screening of dementia. The Wallet Story database included 23 elders with MCI and 12 healthy aging adults, which had 70 narratives in total.

Picture description (PD)

Subjects were asked to look at a picture or a series of pictures that make up a story and describe orally the content in pictures within a limited time. Pictures include the cookie theft (a girl and a boy stealing cookies and a woman washing dishes in the kitchen), the dog story (a boy who hides a dog that he found on the street), the Cinderella story, and so on.

Dementiabank [28] is a multimedia interaction for the study of communication in dementia. Pitt corpus [29], ADReSS database [19], and ADReSSo database [30] are subsets of this database. Pitt corpus mainly included recordings of spoken picture descriptions extracted from participants through the cookie theft picture description from the Boston Diagnostic Aphasia Exam [31], which contained 87 speech recordings in AD patients and 79 speech recordings in healthy controls in the training set, and 71 speech recordings without annotations in the testing set. ADReSS database contained speech samples (WAV format) and transcripts (CHA format) with corresponding MMSE (Mini-Mental State Examination) scores as labels, which included 156 subjects, 108 were for training and 48 were for the test (train:test = 7:3). The ADReSSo database was established after the ADReSS database and included 87 AD patients and 79 HC.

Reading

Transcripts reading

Subjects were given short passages or articles to read aloud and their speeches were recorded. The Gothenburg MCI study was conducted as an experiment with 55 Swedish participants (30 HC and 25 AD) who were instructed by a clinician to read a short passage, consisting of 144 words, as part of their evaluation [32].

Deep learning techniques

In order to investigate the recent progress of deep learning methods in speech-based AD detection, we list some key information in the selected papers in the table below: linguistic tasks, the distribution of participants for each label in the database, the feature types used in papers, the specific model architecture, the model training strategy, and the best performance (Table 2).

Feature types

Feature types mentioned in our paper include demographic features (DeF), duration features (DF), traditional acoustic features (TAF), traditional linguistic features (TLF), acoustic embeddings, and linguistic embeddings. Demographic features include age, years of education, and gender. Duration features contain the duration of the speaker speaking and its statistics. Traditional acoustic features include properties of the sound wave (MFCCs or Formant), speech rate, and the number of pauses. Traditional linguistic features include lexical (word rate or types and their characteristics, e.g., word frequency, repetitions), semantic (word meaning, e.g., idea density), and syntactic (grammar of sentences, e.g., syntactic complexity, grammatical constituents) features. Acoustic embeddings (AE) means the feature vector representations of speech, which can be extracted by ASR models or pretrained models (such as speech BERT or YAMNet). Linguistic embeddings (LE) are a type of automatic feature that refers to the vector representations corresponding to input tokens, which can be obtained by models such as BERT [15], ERNIE [77], or Longformer [78].

Model architectures

In this paragraph, we briefly introduce some deep learning models used in the selected papers, and the model structure used in each paper can be viewed in the table.

Feedforward neural network

Earlier researchers started to use feedforward neural networks (FNN) [79] as feature classifiers in their studies to distinguish healthy people from cognitively impaired patients.

Table 2 Deep learning techniques in all included papers

References	Year	Task	Sample	Feature type	Classifier	Pre-train	Evaluation	Metrics	Best Performance
Bertini, F. et al. [33]	2022	SS-PD-CT1	AD (137)/HC (43)	AE	auDeep	Yes	CV	Accuracy	93.30%
Meghanani, A. et al. [34]	2021	SS-PD-CT2	AD (54)/non-AD (54)	TLF	FNN	No	Test	Accuracy	83.33%
Rohanian, M. et al. [35]	2021	SS-PD-CT3	AD (122)/HC (115)	TAF/TLF/DF	biLSTM	No	Test	Accuracy	84%
Shah Syed, M.S. et al. [36]	2021	SS-PD-CT2	AD (72)/non-AD (72)	TAF	LSTM[a]	No	Test	Accuracy	74.55%
Mahajan, P. et al. [37]	2021	SS-PD-CT2	AD (82)/non-AD (82)	TAF/DF/TLF/DeF	CNN+biLSTM[a]	No	Test	Accuracy	72.92%
Meghanani, A [34].	2021	SS-PD-CT2	AD (78)/non-AD (78)	TAF	CNN+LSTM	No	Test	Accuracy	64.58%
Lindsay, Hali et al. [38]	2021	VF4	HC (66)/MCI (66)	LE	SVM	Yes	CV	AUC	
Rodrigues Makiuchi, M. et al. [39]	2021	SS-PD-CT1SS-CVS1	CT1: AD (168)/HC (98) CVS1: AD (49)/MCI (42)/HC (72)	TAF	GCNN	No	CV	Accuracy	
Liu, Z. et al. [40]	2021	SS-PD-CT1	AD (252)/HC (232)	AE	CNN+biLSTM[a]	Yes	CV	Accuracy	82.59%
Wang, N. et al. [41]	2021	SS-PD-CT3	AD (87)/HC (79)	TLF/LE	C-Attention-Unified model	Yes	Test	Accuracy	80.28%
Bertini, F. et al. [42]	2021	SS-CVS2SS-PD2	eD (16)/MCI (32)/HC (48)	TAF	FNN	Yes	CV	Accuracy	90.57%
Roshanzamir, A. et al. [43]	2021	SS-PD-CT1	AD (170)/HC (99)	LE	LR	Yes	CV	Accuracy	88.08%
Saltz, P. et al. [44]	2021	SS-PD-CT2SS-PD-CT1	CT2: AD (78)/non-AD (78)	LE	BERTXLNet	Yes	CV	Accuracy	
Liu, Z. et al. [45]	2021	SS-PD-CT2	AD (87)/non-AD (79)	TLF/TAF/LE	BERT	Yes	CV	Accuracy	97.18%
Guo, Y. et al. [46]	2021	SS-PD-CT2SS-PD-CT4	CT2: AD (78)/non-AD (78) CT4: AD (115)/HC(839)	LE	BERT	Yes	Test	Accuracy	82.10%
Pan, Y. et al. [47]	2021	SS-PD-CT2	AD (78)/non-AD (78)	LE	BERT large	Yes	Test	Accuracy	84.51%
Chlasta, K. et al. [48]	2021	SS-PD-CT2	AD (78)/non-AD (78)	AE	DemCNN	Yes	Test	Accuracy	62.50%
Gauder, L. et al. [49]	2021	SS-PD-CT2	AD (87)/non-AD (79)	AE	CNN	Yes	Test	Accuracy	78.90%
Haulcy, R. et al. [50]	2021	SS-PD-CT2	AD (78)/non-AD (78)	LE	SVM, RF	Yes	Test	Accuracy	85.40%
Syed, Z.S. et al. [51]	2021	SS-PD-CT2	AD (78)/non-AD (78)	TLF/LE	SVM, LR	Yes	Test	Accuracy	91.67%
Tsai, A.C. Y. et al. [52]	2021	SS-Recall1 & SS-PD-CT6SS-PD-CT1	SS-Recall1 & SS-PD-CT6 : AD (40)/HC (40)CT1: AD (257)/HC (242)	LE	BERT	Yes	Test	Accuracy	
Zhu, Y. et al. [53]	2021	SS-PD-CT2	AD (78)/non-AD (78)	AE/LE	Longformer	Yes	Test	Accuracy	89.58%
Aparna Balagopalan et al. [54]	2021	SS-PD-CT2	AD (78)/non-AD (78)	LE	BERT	Yes	Test	Accuracy	83.32%
Yuan, J. et al. [55]	2021	SS-PD-CT2	AD (78)/non-AD (78)	LE	ERNIE-large	Yes	Test	Accuracy	89.60%

Table 2 (continued)

References	Year	Task	Sample	Feature type	Classifier	Pre-train	Evaluation	Metrics	Best Performance
Xue, C. et al. [56]	2021	SS-CVS9	dementia (330)/MCI (451)/HC (483)	TAF	LSTM	No	CV	Accuracy	67.50%
Roozbeh, S. et al. [57]	2021	SS-PD-CT5	AD (26)/46 (HC)	TAF/TLF	FNN	No	CV	Accuracy	93.05%
Koo, J. et al. [58]	2020	SS-PD-CT2	AD (78)/non-AD (78)	TAF/TLF/AE	CNN+biLSTM[a]	Yes	Test	Accuracy	81.25%
Cummins, N. et al. [59]	2020	SS-PD-CT2	AD (54)/non-AD (54)	TAF/LE	biLSTM[a]	Yes	Test	Accuracy	85.20%
Sarawgi, U. et al. [60]	2020	SS-PD-CT1 SS-PD-CT2	CT1: AD (168)/HC (99)CT2: AD (78)/non-AD (78)	TLF/TAF	FNN	No	CV Test	Accuracy Accuracy	
La Fuente Garcia, S. D. et al. [61]	2020	SS-PD-CT1SS-CVS3	CT1: AD (82)/HC (82) CVS3: AD (30)/HC (16)	TAF	FNN	No	Test	UAR	
Lopez-De-Ipina, K. et al. [62]	2020	VF1	MCI (38)/HC (62)	TAF	CNN	No	CV	Accuracy	92%
Casanova, E. et al. [63]	2020	SS-Recall2SS-Recall3SS-PD4SS-PD5	AD (41)/MCI (55)/HC (194)	TLF	RNN+CRF[a]	Yes	CV	F1-score	81.00%
Pan, Y. et al. [64]	2020	SS-CVS4	ND (21)/MCI (24)/HC (25)	AE	LRSVM	Yes	CV	F1-score	
Searle, T. et al. [65]	2020	SS-PD-CT2	AD (78)/non-AD (78)	LE	DistilBERT	Yes	Test	Accuracy	81%
Li, Y [66].	2020	SS-PD-CT1	AD (155)/HC (145)	DeF/LE/TLF/TLF	LR	Yes	CV	Accuracy	91.25%
Rosas, D.S. et al. [14]	2019	SS-CVS3	Dementia (62)/HC (160)	TLF	FNN	No	Test	Accuracy	86.42%
Chien, Y.W. et al. [67]	2019	SS-Recall5	AD (30)/HC (30)	TAF	biLSTM	Yes	Test	AUC	83.80%
Fritsch, J. et al. [68]	2019	SS-PD-CT1	AD (168)/HC (98)	TLF	LSTM	No	CV	Accuracy	85.60%
Hong, S.Y. et al. [69]	2019	SS-PD-CT1	AD (169)/HC (99)	LE	RNN[a]	Yes	CV	Accuracy	83.50%
Gabor, G. et al. [27]	2019	SS-CVS5SS-Recall4	mAD (25)/MCI (25)/HC (25)	TLF/DF/DeF	SVM	Yes	CV	Accuracy	86.00%
Themistocleous, C. et al. [70]	2018	Reading	HC (30)/MCI (25)	TAF/DeF	FNN	No	CV	Accuracy	83%
Klumpp, P. et al. [71]	2018	SS-PD-CT1	AD (168)/HC (98)	LE	FNN	No	Test	Accuracy	84.40%
Lopez-De-Ipina, K. et al. [72]	2018	VF1SS-CVS6SS-PD3	VF1: MCI (38)/HC (62)CVS6: AD (20)/HC (20) PD3: AD (6)/HC (12)	TAF	CNN	No	CV	Accuracy	
Orimaye, S. O. et al. [73]	2018	SS-PD-CT1	AD task: AD (99)/HC (99) MCI task: MCI (19)/HC (19)	TLF	D2NNLM-5n	No	Test	AUC	
Warnita, T. et al. [74]	2018	SS-PD-CT1	AD (169)/HC (98)	TAF	GCNN	No	CV	Accuracy	73.60%
Chien, Y. W. et al. [23]	2018	VF2VF3	AD (30)/HC (30)	TAF	biLSTM	Yes	Test	AUC	95.40%
Lopez-de-Ipina, K. et al. [12]	2017	VF1SS-CVS6SS-PD3	MCI (40)/HC (60)	TAF	CNN	No	CV	Accuracy	

Table 2 (continued)

References	Year	Task	Sample	Feature type	Classifier	Pre-train	Evaluation	Metrics	Best Performance
Lopez-de-Ipina, K. et al. [21]	2017	VF1	MCI (38)/HC (62)	TAF	CNN	No	CV	Accuracy	75%
D Beltrami et al. [75]	2016	SS-CVS7 SS-PD1	MCI (19)/HC (20)	TLF/TAF	FNN	No	CV CV	F1-score F1-score	
Laszlo, T. et al. [76]	2016	SS-CVS5SS-Recall4	MCI (48)/HC (36)	DF/DeF	SVM	Yes	CV	Accuracy	88.10%
Laszlo, T. et al. [16]	2015	SS-CVS8	MCI (32)/HC (19)	TAF/DF	SVM	Yes	CV	Accuracy	80.40%
Lopez-de-Ipina, K. et al. [26]	2013	SS-CVS6	AD (20)/HC (20)	TAF/DF	FNN	No	CV	Accuracy	94.60%

[a] in Classifier means attention-based method. The full names of abbreviations can be found in "Abbreviations"

Convolution neural network

As the convolutional neural network (CNN) [80] has achieved good results in computer vision tasks, CNN-related models have also begun to be gradually applied to NLP tasks, such as sentence classification, semantic parsing, search query retrieval, and other traditional NLP tasks. Therefore, researchers also began to use CNN models and linear gated convolution neural network (GCNN) [81] to classify speech or text data of AD patients.

Recurrent meural network

In order to add timing information from speaker audio to the model, researchers began to use model architectures including recurrent neural network (RNN) [82], long short-term memory (LSTM) [83], gated recurrent unit(GRU) [84], bidirectional LSTM(BiLSTM) [85], etc. At the same time, researchers also combine these models with CNN or other neural networks, such as pyramidal bidirectional LSTM followed by a CNN layer (pBiLSTM-CNN) proposed by Meghanani. A [86].

Attention-based neural network

With the rise of attention mechanisms [87], researchers began to apply some attention mechanisms to improve the accuracy of the model, such as adding attention mechanisms to RNN models or CNN and LSTM models.

To identify AD with a small amount of data, researchers utilize models pre-trained on large-scale databases as feature extractors to obtain better representations, such as Longformer, BERT, and ERNIE.

Conclusions

What were the characteristics of the databases involved in reported studies?

Twenty-seven different databases were used in 52 studies, in which the appearance frequency of the Pitt corpus and ADReSS database were highest. Fourteen studies used Pitt corpus from Dementiabank, and 19 studies included the ADReSS database.

In 27 databases, 11 languages were used. Twenty-five databases used only one language in one database, including Spain, Chinese, English, Hungarian, Italian, Japanese, Brazilian Portuguese, and Swedish. Two databases used more than one language in one database. For example, AZTIAHO included English, French, Spanish, Catalan, Basque, Chinese, Arabian, and Portuguese.

In 29 databases, labels include AD (Alzheimer's disease), MCI (mild cognitive impairment), and HC (healthy control). Eleven databases contain only AD and HC labels; 7 databases contain only MCI and HC labels; 11 databases contain AD, MCI, and HC labels.

For now, the databases in reported studies were small in single or few languages with uneven distribution. Besides, most were built for cross-sectional studies rather than cohort studies.

What deep learning model architectures were included in reported studies?

Four deep learning methods were applied in these selected papers: FNN, CNN, LSTM, and attention mechanism-based models. Figure 3 shows each number of these methods. These models were generally basic, and embeddings were extracted by models and collected for classification.

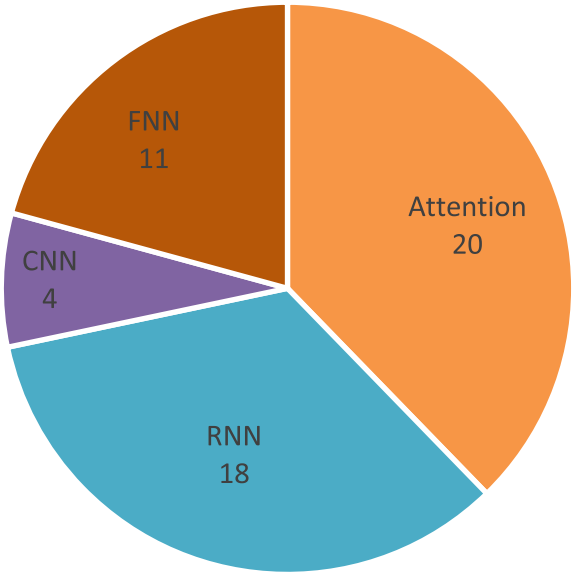

Fig. 3 Paper numbers of FNN, CNN, LSTM, and attention mechanism-based models

How were these deep learning model architectures used in reported studies?

The use of deep learning can be divided into three categories. First, the models trained on the large database were directly used to extract embedding, and then machine learning classifiers were used. Second, the models were pre-trained on a large database and then fine-tuned on dementia-related databases. In some situations, Self-training and data augmentation methods were used in the pre-trained process. Thirdly, deep learning models were built and trained from scratch using dementia-related databases.

What classification performance has been achieved?

The performance advantages of deep learning compared to the traditional method

Balagopalan, A. et al. tested on the ADReSS dataset using different classification models, including SVM, NB, RF, FNN, and BERT. According to the results presented in the paper, when using the FNN method, it can achieve an average accuracy of 77.08% on the ADReSS test set in 3 runs, which is higher than the performance of RF and NB but lower than the average accuracy of 81.25% for the SVM classifier. However, when using BERT, it got the best result for classification with an accuracy of 83.32% [54]. Not only linguistic features, but deep learning has also achieved better results on acoustic features. Bertini, F. et al. used an autoencoder to extract unsupervised features from audio data and then utilized FNN to achieve 93.3% classification accuracy on the Pitt dataset, which is better than the results obtained by traditional machine learning methods such as SVM, NB, and RF [33].

In the detection process of AD, utilizing deep learning methods can effectively improve the performance of the classification models when compared with traditional machine learning methods.

Besides, we compared methods without pre-training and methods with pre-trainig by box plotting in

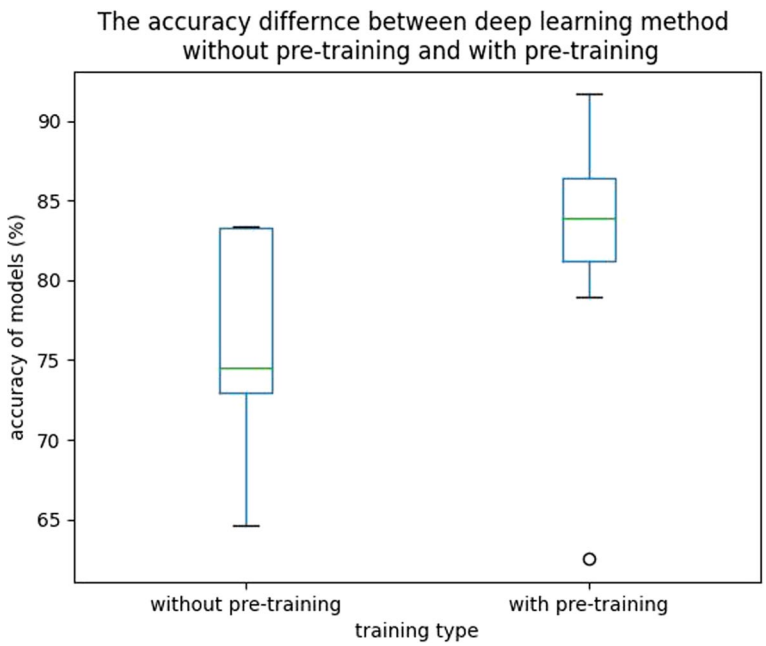

Fig. 4 Comparison of deep learning methods without and with pre-training

SS-PD-CT2 task with a test set for evaluation in Fig. 4. It exhibits that using the pre-training method is more useful than training models from scratch.

Performance difference based on different tasks

On the task selection, SS works better than others tasks generally. In 2017 and 2018, Lopez-De-Ipina, K. et al. conducted research on AD detection based on VF and SS tasks, in which acoustic features were mainly used. The detection accuracies on SS tasks were higher than the result on the VF task [72].

SS tasks can be divided into several different subtasks, including PD, Conversation/interview, and Recall.

In PD tasks, most tasks were based on ADReSS or Pitt database. There were 21 studies that used the ADReSS database and that 11 studies used the Pitt database. The test set on ADReSS database was uniform, detection accuracy in more than 75% of studies can reach more than 80%, and the best result can reach 91.67%. Cross-validation predictions from 85% of studies on the Pitt database exceeded 80% accuracy, and the best result can reach 91.25%. Ten reported studies contain conversation tasks [14, 16, 26, 27, 39, 42, 56, 64, 75, 76].

Though different databases were used, high accuracy can be achieved by cross-validation evaluation, in which 85% of studies exceeded 85% accuracy and the best result can reach 95%.

In Recall tasks, four related studies are included, and all can achieve 80% accuracy.

Comparisons of methods for the ADReSS Challenge

The ADReSS Challenge is the most recent internationally representative speech-based AD detection competition, which was held in Interspeech 2020–2021. The main objective of the ADReSS challenge is to make available a benchmark dataset of spontaneous speech, which is acoustically pre-processed and balanced in terms of age and gender, defining a shared task through which different approaches to AD recognition in spontaneous speech can be compared. Pre-training methods are mainly used in the top five participating teams of the ADReSS challenge, which include two types of useful ways of deep learning techniques.

The first way is pre-training based on deep learning architecture and large datasets, and then fine-tuning on the ADReSS dataset. Saltz, P. et al. [44]; Yuan, J. et al. [55]; and Zhu, Y. et al. [53] used BERT, ERNIE, Longformer-based model architecture to pre-train and then fine-tune, which reached 90%, 89.6%, and 89.58% on ADReSS test set respectively. In terms of characteristics, Saltz, P. et al. and Yuan, J. et al. used linguistic embedding only, and Zhu, Y. et al. used acoustic and linguistic embedding. Besides, Saltz, P. et al. used augmented data during the training stage, Yuan, J. et al. encoded the pause into the transcript and then acquired embedding vector for classification, and Zhu, Y. et al. used Longformer-based transfer learning.

The second way is extracting features based on deep learning architecture, and then training traditional machine learning classifiers based on the extracted features. Syed, Z. S. et al [51] combined traditional linguistic features and linguistic embedding extracted from a pre-trained BERT-based model, and then trained through ensemble learning and fused based on majority-voting, eventually reaching 91.67% accuracy on the ADReSS test set. Haulcy, R. et al. [50] extracted linguistic embedding from BERT with SVM or RF classifier and achieved 85.4% accuracy.

In addition, some other text-based pre-trained models work well. For example, the accuracies of BERT, part of BERT or BERT-based adaptation models [46, 47, 54, 65] were between 81% and 84.51%. Except for the text-based pre-trained models, audio and image-based pre-trained models also have been explored in speech-based AD detection. Chlasta, K. et al [48] trained modified VGGNet architecture to extract acoustic embedding, while Gauder, L. et al. [49] trained wav2vec 2.0 framework to extract acoustic embedding vector, of which both added modified CNN modules for classification, reaching 62.5% and 78.9% accuracy, respectively.

Another training method in the ADReSS Challenge is training from scratch. Traditional linguistic and acoustic features have been applied with the architectures such as FNN [34, 60], attention mechanism-based LSTM [86] and CNN-LSTM [36] model reached 83.33%, 64.58%, and 74.55% accuracy, respectively. After the duration features were added, BiLSTM with highway layers, CNN-BiLSTM-attention-based architecture [35], and dense layer with GRU model [37] reached 84%, 84%, and 72.92% accuracy, respectively.

When using limited clinical data, choosing proper pre-trained task and fine-tuned models are important and effective for disease classification. Generally, CNN-based architectures extract local information, and the LSTM or BERT-based model extracts temporal information. Specifically, pre-training a speech or text encoder with a large speech or text corpus, and using the attention mechanism to map the correspondence, then a fine-tuning model with AD or MCI dataset is a general method to build a framework to train the AD classification from scratch.

The algorithms and performances for detecting MCI

As an intermediate transition state between the normal aging process and mild AD, MCI plays an important role in early screening or AD. Among the screened papers,

16 of them performed MCI detection experiments. 11 of the 16 papers were about distinguishing MCI and healthy people, while the rest were about three classifications of AD, MCI patients, and cognitive normal elders.

For the classification of MCI versus cognitive normal subjects, Lindsay, Hali et al. [38] utilized three different pre-trained models (FastTest, Spacy, Wiki2Vec) to extract word embeddings, then used a SVM classifier to predict labels in different languages (French, German, Dutch), and can achieve 66%, 68%, and 69% AUC, respectively. For three-classification experiments for AD, MCI, and HC, Rodrigues Makiuch, M. et al. [39] using a gated convolutional neural network (GCNN), achieving an accuracy of 60.6% in 40 s of speech data.

MCI manifests as mild cognitive decline. Compared with AD, most MCI patients have less severe memory loss and perform relatively normal on memory tests. As can be seen from the papers we screened, it is more difficult to detect MCI patients than to distinguish AD patients from cognitive normal elders-based speech analysis. And we can find that there are not many studies on MCI detection at present, so it is of great value to further explore the methods of detecting MCI with deep learning techniques.

What were the mainstreams and limitations of reported studies?

The mainstreams and limitations of these selected studies were mainly reflected in language tasks, data modalities, extracted features, and model performance.

Language tasks

Varied databases were built to collect speech from AD and healthy people based on varied tasks. Through the databases we introduced in section 4.2 of this article, we can find that the current mainstream language tasks focus on: Semantic verbal fluency tasks, Spontaneous speech tasks, and some other reading tasks.

Semantic verbal fluency tasks contain animal naming tasks, vegetable, and location naming tasks. As for tasks collecting spontaneous speech, it compromised speech from interviews or conversations speech, recall tasks, and picture description tasks.

From this, we can find that there are many kinds of language tasks, which makes it difficult for researchers to compare their research results.

Therefore, based on the picture description task, the Pitt corpus and the ADReSS database have constructed comparable distribution-balanced databases, and researchers have begun to focus on these two databases for AD classification tests.

However, the languages of Pitt corpus and ADReSS databases are both English, and the amount of data is small, so the current research is also limited to a certain extent.

Data modalities

Based on our table in the "Deep learning techniques" section, we can see that researchers used speech, text, or speech and text to conduct experiments, in which some compared the classification results on the same evaluation test set.

The current research trend is to obtain more characteristic information by combining multimodal data. Different modalities have different representations, so there is some overlap and complementarity of information, as well as a variety of information interactions. Researchers may no longer be limited to the speech and text information of AD patients. Improving the accuracy of the overall decision-making results by integrating multi-modal data such as eye movement data, writing data, and gait performance is also an interesting topic that needs further investigation.

Extracted features

Traditional linguistic and acoustic features were mostly from handcrafted definitions thus these features were explainable. Deep learning-based feature extraction or classification techniques achieved high accuracy for AD classification but short of the lack of interpretability.

Deep learning-based feature extraction methods need a large scale of data, which is hard to precisely define and varies on a different scale of data. Besides, tasks were chosen to pre-train the model for features extraction, for example, ASR or BERT, were not fully compared and analyzed for AD classification tasks.

Model performance

How were these deep learning model architectures used in reported studies? and What classification performance has been achieved? In this paper, the deep learning model architectures and training strategies adopted by the selected papers are presented. In the current study, the researchers use the pre-training model to solve the problem of insufficient training data in AD detection and achieve good results. Most speech-based AD detection using deep learning methods can achieve an accuracy of about 85%. In the ADReSS challenge, some researchers have achieved an accuracy of nearly 90% using pretrained models. However, traditional cognitive impairment screening scales, such as MMSE or MOCA, can usually achieve a screening accuracy of more than 93% [5]. Therefore, as a more convenient AD detection method, speech-based deep learning technology needs to be further improved.

Discussions

Limitation of our studies

In this review, the following limitations may down the outcome confidence level of our paper:

(1) In the process of paper search, our search keywords are missing "pre-trained model," which leave out some papers that refer to "pre-trained model" but do not mention "deep learning" or "neural network". Although we add some papers from other sources, this problem increases the risk of bias of the paper search results.
(2) Because of our selection criteria, only papers written in English were selected, which resulted in some non-English databases and studies not being included in this review, thus increasing the language bias and affecting some language-related features.
(3) Due to the overlap of deep learning methods in many papers, for example, the classifier proposed by Liu, Z. et al. is a combination of CNN, BiLSTM, and attention, so it is difficult to separate it into a specific deep learning category [40]. The lack of a very clear standard in the process of classifying deep learning methods also increases the error of statistical analysis to a certain extent.
(4) In the process of analyzing the performance of deep learning models, there may be some potential risks of bias. Because we were only focused on the best performance of the model in the paper, different databases, different testing methods, and different evaluation indicators may possibly lead to a skewed understanding that how well the algorithms worked.

Research directions

The purpose of this review paper is to investigate current researchers' application of deep learning methods for speech-based AD detection and to explore future possibilities. The current dementia-related databases are usually small, with a single language, uneven distribution, and inconsistent tasks. However, fusing the multi-modal data rather than using only one modality can extract more useful information for the classification of AD patients, and the application of pre-trained models can also greatly improve the classification accuracy. Another point to note is that the databases in the papers we screened lack cohort study data, so it is difficult to prove the reliability of the results of speech analysis on intra-individual repeated testing. Besides, currently, speech-based AD detection has not been widely applied clinically.

So our future goals are as follows:

(1) To establish and publish a balance-distributed Chinese AD database, including the speech data of the picture-distribution task and the writing data of the clock-drawing test.

At the same time, we hope researches can collect cohort data to study the tracking performance of speech analysis in individual patients over time.

(2) To explore the potential of new deep learning models to improve classification accuracy by utilizing speech, writing, and other multi-modal data.

Improving the interpretability of feature representations that have been extracted by deep learning methods in the assessment of cognitive impairment.

(3) To establish efficient and accurate computer-aided diagnosis methods, which can shorten the time of large-scale AD screening. The study on AD detection also promotes the development of portable diagnostic devices, which could timely detect AD and timely intervene to delay the disease.
(4) In addition to Alzheimer's disease, there are other causes of dementia, so we hope that future researchers can use speech analysis to detect other types of dementia.

Abbreviations

AD: Alzheimer's disease; MCI: Mild cognitive impairment; HC: Healthy control; ND: Neurodegenerative disorders; mAD: Mild Alzheimer's disease; eD: Early dementia; MRI: Resonance imaging measurement; PGA: Gipuzkoa-Alzheimer Project; SVF: Semantic verbal fluency; SS: Spontaneous speech; PD: Picture description; TLF: Traditional linguistic features; TAF: Traditional acoustic features; DeF: Demographic features; DF: Duration features; LE: Linguistic embeddings; AE: Acoustic embeddings; CRF: Conditional random field; CV: Cross-validation; UAR: Unweighted average recall; AUC: Area under curve; DNN: Deep neural network; GMM: Gaussian Mixture Model; FNN: Feedforward neural network; SVM: Support vector machine; LR: Logistic regression; RF: Random Forest; ASR: Automatic speech recognition; NLP: Natural language processing; BERT: Bidirectional Encoder Representations from Transformers; ERNIE: Enhanced Representation through kNowledge IntEgration; CNN: Convolutional neural network; GCNN: Gated Convolution Neural Network; GRU: Gated recurrent unit; LSTM: Long short-term memory; RNN: Recurrent neural network; BiLSTM: Bidirectional LSTM; DemCNN: Custom audio convolutional neural network; pBiLSTM-CNN: Pyramidal bidirectional LSTM followed by a CNN layer; D2NNLM-5n: Deep neural network and deep language models with decomposed 5-gram feature.

Acknowledgements

Not applicable

Authors' contributions

The manuscript was designed by Qin Yang and Xinyun Ding. Qin Yang conducted a literature review and summary. Qin Yang and Xinyun Ding drafted the manuscript. Feiyang Xu, Xin Li, and Zhenhua Ling edited and revised the manuscript. All authors read and approved the final manuscript.

Funding
This work was partially funded by the National Nature Science Foundation of China (Grant No. 62106246) and the China Postdoctoral Science Foundation (Grant No. 2021M693101).

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]NELSLIP, University of Science and Technology of China, Hefei, China. [2]iFlytek Research, iFlytek Co.Ltd, Hefei, China.

Received: 23 March 2022 Accepted: 23 November 2022
Published online: 14 December 2022

References

1. Monica Moore MSG, Díaz-Santos M, Vossel K. Alzheimer's Association 2021 Facts and Figures Report[J].
2. Morley JE, Morris JC, Berg-Weger M, Borson S, Carpenter BD, Del Campo N, et al. Brain health: The importance of recognizing cognitive impairment: An iagg consensus conference. J Am Med Dir Assoc. 2015;16:731–9 Elsevier.
3. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimers Dement. 2011;7:263–9 Elsevier.
4. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify alzheimer's disease in narrative speech. J Alzheimers Dis. 2016;49:407–22 IOS Press.
5. Satt A, Hoory R, König A, Aalten P, Robert PH. Speech-based automatic and robust detection of very early dementia. Fifteenth annual conference of the international speech communication association. 2014.
6. Hoffmann I, Nemeth D, Dye CD, Pákáski M, Irinyi T, Kálmán J. Temporal parameters of spontaneous speech in alzheimer's disease. Int J Speech Lang Pathol. 2010;12:29–34 Taylor & Francis.
7. Croisile B, Brabant M-J, Carmoi T, Lepage Y, Aimard G, Trillet M. Comparison between oral and written spelling in alzheimer's disease. Brain Lang. 1996;54:361–87 Elsevier.
8. Croisile B, Ska B, Brabant M-J, Duchene A, Lepage Y, Aimard G, et al. Comparative study of oral and written picture description in patients with alzheimer's disease. Brain Lang. 1996;53:1–19 Elsevier.
9. Cuetos F, Arango-Lasprilla JC, Uribe C, Valencia C, Lopera F. Linguistic changes in verbal expression: A preclinical marker of alzheimer's disease. J Int Neuropsychol Soc. 2007;13:433–9 Cambridge University Press.
10. Markaki M, Stylianou Y. Voice pathology detection and discrimination based on modulation spectral features. IEEE Trans Audio Speech Lang Process. 2011;19:1938–48.
11. Yang Q, Xu F, Ling Z, et al. Selecting and Analyzing Speech Features for the Screening of Mild Cognitive Impairment[C]//2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021:1906-1910.
12. de Lizarduy UM, Salomón PC, Vilda PG, et al. ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment[J]. Loquens. 2017;4(1):e037-e037.
13. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A-r, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag. 2012;29:82–97 IEEE.
14. Rosas DS, Arriaga ST, Fernandez MAA. Search for dementia patterns in transcribed conversations using natural language processing. 2019 16th international conference on electrical engineering, computing science and automatic control, cce. 2019. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85075099563\&doi=10.1109%2fICEEE.2019.8884572\&partnerID=40\&md5=7440614079b3a790ea15b823c4265d76 https://ieeexplore.ieee.org/document/8884572/.
15. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
16. Tóth L, Gosztolya G, Vincze V, et al. Automatic detection of mild cognitive impairment from spontaneous speech using ASR[C]. ISCA, 2015.
17. Vigo I, Coelho L, Reis S. Speech- and language-based classification of alzheimer's disease: A systematic review. Bioengineering (Basel). 2022;9 Available from: https://www.ncbi.nlm.nih.gov/pubmed/35049736.
18. Petti U, Baker S, Korhonen A. A systematic literature review of automatic alzheimer's disease detection from speech and language. J Am Med Inform Assoc. 2020;27:1784–97 Oxford University Press.
19. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. arXiv preprint arXiv:200406833. 2020.
20. Lopes M, Brucki SMD, Giampaoli V, Mansur LL. Semantic verbal fluency test in dementia: Preliminary retrospective analysis. Dement Neuropsychol. 2009;3:315–20.
21. Lopez-De-Ipina K, Martinez-De-Lizarduy U, Calvo PM, Beitia B, Garcia-Melero J, Ecay-Torres M, et al. Analysis of disfluencies for automatic detection of mild cognitive impartment: A deep learning approach. 2017 international work conference on bio-inspired intelligence: Intelligent systems for biodiversity conservation, iwobi 2017 - proceedings. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028563436\&doi=10.1109%2fIWOBI.2017.7985526\&partnerID=40\&md5=619c2b86a411eac9a4a849fbb9063ba5.
22. Campagna F, Montagnese S, Ridola L, Senzolo M, Schiff S, De Rui M, et al. The animal naming test: An easy tool for the assessment of hepatic encephalopathy. Hepatology. 2017;66:198–208.
23. Chien YW, Hong SY, Cheah WT, et al. An Assessment System for Alzheimer's Disease Based on Speech Using a Novel Feature Sequence Design and Recurrent Neural Network[C]//2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE. 2018:3289-3294.
24. MacWhinney B, Fromm D, Forbes M, Holland A. AphasiaBank: Methods for studying discourse. Aphasiology. 2011;25:1286–307.
25. Illes J. Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, huntington's, and parkinson's. Brain Lang. 1989;37:628–42.
26. López-de-Ipiña K, Alonso JB, Travieso CM, et al. On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis[J]. Sensors. 2013;13(5):6730-6745.
27. Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. Comput Speech Lang. 2019;53:181–97.
28. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology. 1994;51(6):585-594.
29. Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL. The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. Arch Neurol. 1994;51:585–94 American Medical Association.
30. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: The adresso challenge. arXiv preprint arXiv:210409356. 2021;
31. Goodglass H, Kaplan E, Weintraub S. BDAE: The boston diagnostic aphasia examination. Philadelphia: Lippincott Williams & Wilkins; 2001.
32. Graves WW, Desai R, Humphries C, Seidenberg MS, Binder JR. Neural systems for reading aloud: A multiparametric approach. Cereb Cortex. 2010;20:1799–815.
33. Bertini F, Allevi D, Lutero G, et al. An automatic Alzheimer's disease classifier based on spontaneous spoken English[J]. Computer Speech & Language. 2022;72:101298.
34. Meghanani A, Anoop CS, Ramakrishnan AG. Recognition of alzheimer's dementia from the transcriptions of spontaneous speech using fastText and cnn models. Front Comput Sci. 2021;3 Available from: https://www.

scopus.com/inward/record.uri?eid=2-s2.0-85117879671\&doi=10.3389%2ffcomp.2021.624558\&partnerID=40\&md5=8802a1bb3591d7ac3ae4427d565ff826.
35. Rohanian M, Hough J, Purver M. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. Proceedings of the annual conference of the international speech communication association, interspeech. p. 4191–5. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117836710\&doi=10.21437%2fInterspeech.2021-1633\&partnerID=40\&md5=3ea83de2cd6059a2b07e9673c1fa8ad5.
36. Shah Syed MS, Syed ZS, Pirogova E, Lech M. Static vs. dynamic modelling of acoustic speech features for detection of dementia. Int J Adv Comput Sci Appl. 2020;11:662–7 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101485611\&doi=10.14569%2fIJACSA.2020.0111082\&partnerID=40\&md5=6012b39cddf348bffb633d0cec4a10dc https://thesai.org/Downloads/Volume11No10/Paper_82-Static_vs_Dynamic_Modelling_of_Acoustic_Speech_Features.pdf.
37. Mahajan P, Baths V. Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. Front Aging Neurosci. 2021;13 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101241179\&doi=10.3389%2ffnagi.2021.623607\&partnerID=40\&md5=5adf0b6ee0702b74fce0d978b39fc46e.
38. Lindsay H, Müller P, Kröger I, et al. Multilingual Learning for Mild Cognitive Impairment Screening from a Clinical Speech Task[C]//Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). 2021:830-838.
39. Rodrigues Makiuchi M, Warnita T, Inoue N, Shinoda K, Yoshimura M, Kitazawa M, et al. Speech paralinguistic approach for detecting dementia using gated convolutional neural network. IEICE Trans Inf Syst. 2021;104:1930–40 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119404517\&doi=10.1587%2fTRANSINF.2020EDP7196\&partnerID=40\&md5=5724ad7f872c34ee3dd22594134bfe2f https://www.jstage.jst.go.jp/article/transinf/E104.D/11/E104.D_2020EDP7196/_pdf/-char/en.
40. Liu Z, Guo Z, Ling Z, Li Y. Detecting alzheimer's disease from speech using neural networks with bottleneck features and data augmentation. ICASSP, ieee international conference on acoustics, speech and signal processing - proceedings. p. 7323–7. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115045359\&doi=10.1109%2fICASSP39728.2021.9413566\&partnerID=40\&md5=2de803012c685385d66cb95905705bd1 https://ieeexplore.ieee.org/document/9413566/.
41. Wang N, Cao Y, Hao S, Shao Z, Subbalakshmi KP. Modular multi-modal attention network for alzheimer's disease detection using patient audio and language data. Proceedings of the annual conference of the international speech communication association, interspeech. p. 4196–200. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119294157\&doi=10.21437%2fInterspeech.2021-2024\&partnerID=40\&md5=9097007b8625a261b30ff9f4ffd91e63.
42. Bertini F, Allevi D, Lutero G, Montesi D, Calzà L. Automatic speech classifier for mild cognitive impairment and early dementia. 2021;3:Article 8. Available from: https://doi.org/10.1145/3469089.
43. Roshanzamir A, Aghajan H, Soleymani BM. Transformer-based deep neural network language models for alzheimer's disease risk assessment from targeted speech. BMC Med Inform Decis Mak. 2021;21. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102362746\&doi=10.1186%2fs12911-021-01456-3\&partnerID=40\&md5=95521a68019b47b578ee41d2eb335b00. https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-021-01456-3.pdf.
44. Saltz P, Lin SY, Cheng SC, Si D. Dementia detection using transformer-based deep learning and natural language processing models, Proceedings - 2021 ieee 9th international conference on healthcare informatics, ischi; 2021. p. 509–10. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118099471\&doi=10.1109%2fICHI52183.2021.00094\&partnerID=40\&md5=151a304b9e65a6559a5f487e625d36b1 https://ieeexplore.ieee.org/document/9565750/
45. Liu Z, Proctor L, Collier PN, Zhao X. Automatic diagnosis and prediction of cognitive decline associated with alzheimer's dementia through spontaneous speech. 2021 ieee international conference on signal and image processing applications (icsipa). p. 39–43. Available from: https://ieeexplore.ieee.org/document/9576784/.
46. Guo Y, Li C, Roan C, Pakhomov S, Cohen T. Crossing the "cookie theft" corpus chasm: Applying what bert learns from outside data to the adress challenge dementia detection task. Front Comput Sci. 2021;3 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117882916\&doi=10.3389%2ffcomp.2021.642517\&partnerID=40\&md5=50a9d3ba79b81786ad23d1c42abdafee.
47. Pan Y, Mirheidari B, Harris JM, Thompson JC, Jones M, Snowden JS, et al. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. Proceedings of the annual conference of the international speech communication association, interspeech. p. 4216–20. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117802392\&doi=10.21437%2fInterspeech.2021-1519\&partnerID=40\&md5=9896ec0dec5e6f6c6c0b2386ea8cee9a.
48. Chlasta K, Wołk K. Towards computer-based automated screening of dementia through spontaneous speech[J]. Frontiers in Psychology. 2021;11:623237.
49. Gauder L, Pepino L, Ferrer L, Riera P. Alzheimer disease recognition using speech-based embeddings from pre-trained models. Proceedings of the annual conference of the international speech communication association, interspeech. p. 4186–90. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117821689\&doi=10.21437%2fInterspeech.2021-753\&partnerID=40\&md5=935b981da22de50b19239b345c1e4886.
50. Haulcy R, Glass J. Classifying alzheimer's disease using audio and text-based representations of speech. Front Psychol. 2021;11 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100117187\&doi=10.3389%2ffpsyg.2020.624137\&partnerID=40\&md5=0285ad5a59684ba3147c3c5cb543f9b3.
51. Syed ZS, Syed MSS, Lech M, et al. Automated recognition of Alzheimer's dementia using bag-of-deep-features and model ensembling[J]. IEEE Access. 2021;9:88377-88390.
52. Tsai ACY, Hong SY, Yao LH, Chang WD, Fu LC, Chang YL. An efficient context-aware screening system for alzheimer's disease based on neuropsychology test. Sci Rep. 2021;11. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115436113\&doi=10.1038%2fs41598-021-97642-4\&partnerID=40\&md5=74a148b1c8a1b57173f133b5c6479281. https://www.nature.com/articles/s41598-021-97642-4.pdf.
53. Zhu Y, Liang X, Batsis JA, Roth RM. Exploring deep transfer learning techniques for alzheimer's dementia detection. Front Comput Sci. 2021;3 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117919556\&doi=10.3389%2ffcomp.2021.624683\&partnerID=40\&md5=096a607aff4ca369ec2676ac21c360ee.
54. Balagopalan A, Eyre B, Robin J, Rudzicz F, Novikova J. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. Front Aging Neurosci. 2021;13:189.
55. Yuan J, Cai X, Bian Y, et al. Pauses for detection of Alzheimer's disease[J]. Frontiers in Computer Science. 2021;2:624488.
56. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: A framingham heart study. Alzheimers Res Ther. 2021;13. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114013462\&doi=10.1186%2fs13195-021-00888-3\&partnerID=40\&md5=8008fb2529fc37ef399baabdea4fe895. https://alzres.biomedcentral.com/track/pdf/10.1186/s13195-021-00888-3.pdf.
57. Sadeghian R, Schaffer JD, Zahorian SA. Towards an Automatic Speech-Based Diagnostic Test for Alzheimer's Disease[J]. Frontiers in Computer Science. 2021;3:624594.
58. Koo J, Lee JH, Pyo J, Jo Y, Lee K. Exploiting multi-modal features from pre-trained networks for alzheimer's dementia recognition. Proceedings of the annual conference of the international speech communication association, interspeech. p. 2217–21. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098223185\&doi=10.21437%2fInterspeech.2020-3153\&partnerID=40\&md5=9749b41613248885ab47dd0b6eb49019.
59. Cummins N, Pan Y, Ren Z, Fritsch J, Nallanthighal VS, Christensen H, et al. A comparison of acoustic and linguistics methodologies for alzheimer's dementia recognition. Proceedings of the annual conference of the international speech communication association, interspeech. p. 2182–6. Available from: https://www.scopus.com/inward/record.uri?

eid=2-s2.0-85098104245\&doi=10.21437%2fInterspeech.2020-2635\&partnerID=40\&md5=279b53631756260e173505b17c8f7b3c.
60. Sarawgi U, Zulfikar W, Soliman N, Maes P. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. Proceedings of the annual conference of the international speech communication association, interspeech. p. 2212–6. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098161068\&doi=10.21437%2fInterspeech.2020-3137\&partnerID=40\&md5=8b619bdc30a02a24448fdc721e2ec709.
61. La Fuente Garcia SD, Haider F, Luz S. Cross-corpus feature learning between spontaneous monologue and dialogue for automatic classification of alzheimer's dementia speech. Proceedings of the annual international conference of the ieee engineering in medicine and biology society, embs. p. 5851–5. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091007128\&doi=10.1109%2fEMBC44109.2020.9176305\&partnerID=40\&md5=3d05d67cf6620793cd1811a54a272d77 https://ieeexplore.ieee.org/document/9176305/.
62. López-de-Ipiña K, Martinez-de-Lizarduy U, Calvo PM, Beitia B, García-Melero J, Fernández E, et al. On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment. Neural Comput & Applic. 2020;32:15761–9 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85046659670\&doi=10.1007%2fs00521-018-3494-1\&partnerID=40\&md5=1cfd5f2901597f44e51889ffe5fbf5eb.
63. Casanova E, Treviso MV, Hübner LC, Aluísio SM. Evaluating sentence segmentation in different datasets of neuropsychological language tests in brazilian portuguese. LREC 2020 - 12th international conference on language resources and evaluation, conference proceedings. p. 2605–14. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096526896\&partnerID=40\&md5=a1047325eae3924ba99289c9b88537cc.
64. Pan Y, Mirheidari B, Tu Z, O'Malley R, Walker T, Venneri A, et al. Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification. Proceedings of the annual conference of the international speech communication association, interspeech. p. 4806–10. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098228270\&doi=10.21437%2fInterspeech.2020-2684\&partnerID=40\&md5=c2e88eba4a60ff4ecaa6f6bc0973fe4e.
65. Searle T, Ibrahim Z, Dobson R. Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. Proceedings of the annual conference of the international speech communication association, interspeech. p. 2192–6. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098132719\&doi=10.21437%2fInterspeech.2020-2729\&partnerID=40\&md5=65e5a73cbbef50ea82abdf47051356a2.
66. 2020. p. Article 65. Available from: https://doi.org/10.1145/3446132.3446197.
67. Chien YW, Hong SY, Cheah WT, et al. An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network[J]. Scientific Reports. 2019;9(1):1-10.
68. Fritsch J, Wankerl S, Noth E. Automatic diagnosis of alzheimer's disease using neural network language models. ICASSP, ieee international conference on acoustics, speech and signal processing - proceedings. p. 5841–5. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85069004064\&doi=10.1109%2flCASSP.2019.8682690\&partnerID=40\&md5=a76b08675c5fc83207db7f93eb40deb0 https://ieeexplore.ieee.org/document/8682690/.
69. Hong SY, Yao LH, Cheah WT, Chang WD, Fu LC, Chang YL. A novel screening system for alzheimer's disease based on speech transcripts using neural network. Conference proceedings - ieee international conference on systems, man and cybernetics. p. 2440–5. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076790074\&doi=10.1109%2fSMC.2019.8914628\&partnerID=40\&md5=0ba33c950839d96612ac11b418c1b7bf https://ieeexplore.ieee.org/document/8914628/.
70. Themistocleous C, Eckerström M, Kokkinakis D. Identification of mild cognitive impairment from speech in swedish using deep sequential neural networks. Front Neurol. 2018;9 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078279354\&doi=10.3389%2ffneur.2018.00975\&partnerID=40\&md5=4e70dc71cfe974ba2f6b83d759cc36a0.
71. Klumpp P, Fritsch J, Nöth E. ANN-based alzheimer's disease classification from bag of words. Speech communication - 13th itg-fachtagung sprachkommunikation. p. 341–4. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85068984448\&partnerID=40\&md5=cdd22c15aea57b1b7a80e7aef8fed859.
72. López-De-Ipiña K, Martinez-De-Lizarduy U, Calvo PM, Mekyska J, Beitia B, Barroso N, et al. Advances on automatic speech analysis for early detection of alzheimer disease: A non-linear multi-task approach. Curr Alzheimer Res. 2018;15:139–48 Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042557141\&doi=10.2174%2f1567205014666171120143800\&partnerID=40\&md5=ef13eda656195e18478270a3578f642d https://www.eurekaselect.com/article/86986.
73. Orimaye SO, Wong JSM, Wong CP. Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. PLoS ONE. 2018;13. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85056276091\&doi=10.1371%2fjournal.pone.0205636\&partnerID=40\&md5=8f4406472880338ca680fcb16f54d4eb. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0205636\&type=printable.
74. Warnita T, Inoue N, Shinoda K. Detecting alzheimer's disease using gated convolutional neural network from audio data. Proceedings of the annual conference of the international speech communication association, interspeech. p. 1706–10. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055009906\&doi=10.21437%2fInterspeech.2018-1713\&partnerID=40\&md5=b54c660f82a0c8e14de76bbaf5ec3177.
75. Beltrami D, Calzà L, Gagliardi G, et al. Automatic identification of mild cognitive impairment through the analysis of Italian spontaneous speech productions[C]//Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2016:2086-2093.
76. Gosztolya G, Tóth L, Grósz T, Vincze V, Hoffmann I, Szatlóczki G, et al. Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. Proc Interspeech. 2016;2016:107–11.
77. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:190507129. 2019.
78. Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv preprint arXiv:200405150. 2020.
79. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6 Nature Publishing Group.
80. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM. 2017;60(6):84-90.
81. Najibi M, Rastegari M, Davis LS. G-cnn: An iterative grid based object detector. Proceedings of the ieee conference on computer vision and pattern recognition. 2016. p. 2369–77.
82. Elman JL. Finding structure in time. Cogn Sci. 1990;14:179–211 Wiley Online Library.
83. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9:1735–80 MIT Press.
84. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014.
85. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 1997;45:2673–81 Ieee.
86. Meghanani A, C. S A, Ramakrishnan AG. An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech. 2021 ieee spoken language technology workshop, slt 2021 - proceedings. p. 670–7. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103974214\&doi=10.1109%2fSLT48900.2021.9383491\&partnerID=40\&md5=d20928366bdbad7e806d99fa9a073bc4 https://ieeexplore.ieee.org/document/9383491/.
87. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems. 2017;30.

Publisher's Note